DOCUMENT RESUME

ED 253 562 TM 850 102

AUTHOR Slotnick, Henry B.
TITLE Instructional Development Ways and Means: Test Scoring and Analysis: Norm-Referenced, Criterion-Referenced, and Domain-Referenced Tests.
SPONS AGENCY North Dakota Univ., Grand Forks. Office of Instructional Development.
PUB DATE 84
NOTE 38p.; For related documents, see ED 224 423-427.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Answer Sheets; *Computer Software; *Criterion Referenced Tests; Educational Testing; Higher Education; Item Analysis; Medical Education; *Norm Referenced Tests; *Scoring; Teacher Made Tests; Test Construction; Test Theory
IDENTIFIERS *Domain Referenced Tests; University of North Dakota

ABSTRACT

This manual is designed to assist faculty members at the University of North Dakota (UND) in the construction, scoring, and analysis of their classroom tests. A computer program is described which will assist staff in test scoring and analysis. The program was developed by the author in the Office of Medical Education and Evaluation of the University of North Dakota. It is accessible at the UND Computer Center. The program is capable of scoring and analyzing norm-referenced, criterion-referenced, and domain-referenced tests. The manual describes the three types of instruments and suggests appropriate uses for each. It also provides information on the preparation of scoring keys and interpretation of test results. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED255562

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
[X] This document has been reproduced as received from the person or organization originating it.
[ ] Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

INSTRUCTIONAL DEVELOPMENT WAYS AND MEANS

Test Scoring and Analysis: Norm-Referenced, Criterion-Referenced, and Domain-Referenced Tests

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. E. Young
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Henry B. Slotnick
Office of Medical Education & Evaluation
University of North Dakota

A joint project of the
Office of Medical Education & Evaluation,
Office of Instructional Development,
and Computer Center

Requests for copies can be made
to any of these offices

Copyright
University of North Dakota
1984

Test Scoring and Analysis:
Norm-Referenced, Criterion-Referenced,
and Domain-Referenced Tests

This manual describes a program, accessible at the UND Computer Center, which will assist faculty members in scoring and analyzing their classroom tests. The program has been developed by Henry B. Slotnick (Office of Medical Education & Evaluation). Assistance with this program, and with the construction of effective tests of student achievement, is provided by the Computer Center (computer-related assistance) and the Offices of Medical Education & Evaluation and Instructional Development (test development and interpretation).

1.0 Introduction This program is designed for a variety of uses including the scoring of tests, the analysis of tests and items for their effectiveness, and, under certain circumstances, the calculating of grades. The program generates information about individual students, the class as a whole, individual test questions (called items), and the total test. Also, the program has some unique features: It scores subtests (subparts of the test) as well as the total test; it allows test items to be included in different subtests; it scores items with more than one correct answer and allows partial credit for responses. Following the directions in this manual, only a minimal number of decisions need be made in order to score a test and receive information on both the test and the students. In all but one case, user decisions are reported to the computer on test answer keys--the first sheets read by an optical scanner. And, although the program is easy to use, help is available to anyone needing additional information or aid in interpreting results. Faculty members can call the Office of Medical Education (3833) of the Office of Instructional Development (3325).

1.1 Kinds of Tests Scored The most useful feature of this program is its capability to score and analyze different kinds of tests: norm-referenced, criterion-referenced, and domain-referenced. As used here, a test is defined as

> a forced-choice examination designed to provide information about a student's status relative to some body of knowledge or skill.

Where: Forced choice questions require the student to select one answer from among a variety of options provided by the test constructor; the student recognizes (rather than provides) the correct response. Commonly used forced-choice formats include true/false, multiple choice, and matching.

Body of knowledge of skill is the area covered by the test; the test questions operationally define what a student is to know or do.

> Status is simply the way of describing how much the student knows or can do. Status can be described in a variety of ways (e.g., in comparison to the performance of other students, relative to an absolute standard, as a percent of the items attempted).

Indeed, the method of reporting status is an attribute of the kind of test the faculty member is using. If the student's status is described relative to the performances of others, the test is likely Norm-Referenced (NRT)*, while a Criterion-Referenced Test (CRT) means the student's core is compared to an absolute standard. A Domain-Referenced Test (DRT) reports the percent of the domain (the body of knowledge or skill the student is to have learned) the student has mastered.

1.2 Use of This Manual A user will make certain decisions about how their tests will be scored, set up the scoring keys, and interpret the program's output. Information describing these activities is covered in this manual. Users are STRONGLY ENCOURAGED to read this manual, and to refer to the index and glossary for other information on an as-needed basis.

1.3 Futher References This manual covers the scoring and analysis of tests. A companion Instructional Development Ways and Means, titled Grade Book: Computer-Assisted Record Keeping, offers an efficient way to record and calculate students' grades. Also, Testing and Grading, a third publication, provides useful guidance in developing effective tests of college level learning. Section 7.4 of this manual presents a short checklist for planning a test and constructing test questions. Additional information and assistance regarding testing and grading is available from both the Office of Medical Education & Evaluation and the Office of Instructional Development.

1.4 Answer Sheets Only NCS General Purpose Answer Sheets can be read by the UND scanner. These sheets are available for purchase through the University Bookstore.

---

*The three kinds of tests--NRT, CRT, and DRT--are defined in terms of additional attributes, beyond the manner in which the student's status is reported. See these terms in the glossary for a more complete definition of the tests' attributes.

2.0 Table of Contents

Section		Page
1.0	Introduction	1
1.1	Kinds of Tests Scored	1
1.2	Use of This Manual	2
1.3	Further References	2
1.4	Answer Sheets	2
2.0	Table of Contents	3
3.0	General Considerations	4
3.1	Answer Keys	4
3.2	Student Answer Sheets	7
3.3	Submitting Answer Sheets to be Scored	7
3.4	Printout and Its Interpretation	7
4.0	Norm-Referenced Tests	10
4.1	Good and Poor Uses of NRT's	10
4.2	Preparing Scoring Keys	11
4.3	Interpreting Results	11
5.0	Criterion-Referenced Tests	12
5.1	Good and Poor Uses of CRT's	12
5.2	Preparing Scoring Keys	13
5.3	Interpreting Results	13
6.0	Domain-Referenced Tests	14
6.1	Good and Poor Uses of DRT's	14
6.2	Preparing Scoring Keys	14
6.3	Interpreting Results	14
7.0	Addenda	15
7.1	Glossary	15
7.2	Examples	17
7.3	Planning a Test	19
7.4	Rules of Thumb: Essay, Objective & Performance Tests	20

3.0 General Considerations The four topics in this section (preparing answer keys, student answer sheets, submitting tests to be scored, and interpreting output) are common to all test types. Consequently, this section of the manual must be read by all users.

3.1 Answer Keys The first sheets read by the scanner are the answer keys - the sheets that provide the computer with the information it needs to score each student's test and tabulate information on each item in the examination. In all cases, these sheets have the word "KEY" in columns 18-20 of the answer sheet (see the examples, section 7.3).

Four kinds of information are required: The test name (which is used to label the printouts produced), the kind of test being scored (NRT, CRT, or DRT), the number of points awarded for each question, and subscale information (the name of each subtest, the number of subtests, and the items appearing in each). Each kind of information appears on the key for the test.

3.1.1 Main Key This is the first sheet read by the scanner. It contains the following types of information in the specified locations (again see section 7.3).

Information	Location	Explanation
Test Name	Columns 1-14 (in the "name block: on the answer sheet)	This is the name assigned to the test by the user. The name is used to label the printout only.
Test Type	Columns 15-17	Specify "NRT", "CRT", or "DRT" according to how student status is to be reported. See sections 1.1 and the glossary for descriptions of each type. Default is NRT*.
"KEY"	Columns 18-20	This word MUST appear on each answer key sheet in the columns indicated. "KEY" is not included on students' answer sheets.
Number of questions (items) on the test	Columns A-C (in the identification block)	This number should be right justified. If the field is left blank, the computer will automatically count the number of items on the test. Note that this is the number of items, and not the total point value of the items.

*See definition of Default in the Glossary.

Information	Location	Explanation
Number of Criteria only if the test is a CRT (leave blank for NRT's and DRT's)	Column E	There is a maximum of 5 criteria for any CRT, and these refer to the total testscore. Subscales are not scored using the criteria.
Number of Subscales	Column F	This is the number of subparts of the test to be scored and reported. Additional information on subscales will be provided on other answer keys. The maximum is 9 and the default is 0 (meaning the test is not divided into subparts).
Number of different point values	Column H	Point values are the numbers of points which can be awarded for selecting particular options. For example, if 5 points were awarded for correct response and 3 points for partially correct answers, there would be a 2 in column H; the computer would then have to be told which options received 5 points each and which 3 points.
Point value for options on this key	Columns I-J	The number of points awarded for the items blackened on the main key. For instance, if there were 1 point for each correct answer, column I would be left blank, and J would have a "1" in it. Then, blackening in the correct option in the spaces provided on the balance of the answer sheet means that every option identified was worth 1 point.
Test questions	Items 1-200	Blacken one option per item receiving the number of points indicated in columns I-J.

3.1.2 Subkey Sheets The number of subkey sheets is one less than the number of different point values in column H on the main key. If two points are awarded for a correct response, and one point for a partially correct answer, there would be one subkey sheet; the 2-point options would be indicated on the main key, and the 1-point options on the single subkey sheet. Similarly, if there is only a single point value for correct answers (e.g., students receive 2 points for correct responses, 0 for all others), 2-point options would be indicated on the main key and there would be no subkey sheets at all.

Subkey sheets follow the main key. Each has the point value in columns I-J, and the options receiving those point values are blackened in. Aside from columns 18-20 (with the word "KEY"), columns I-J are the only blackened in on the subkey sheets.

Sometimes, two answers are correct for a question, and this problem is handled with subkey sheets. If options A and B for questions 17 are both correct (and 1 point is awarded for correct answers), a "2" would be coded in column H on the main key (the number of different point values), and a "1" in column J. Option A for item 17 would be blackened on that sheet (along with the correct responses for the other questions on the test), and a subkey sheet would be prepared with a "1" in column J and "B" blackened in for item 17. This would allow the computer to award 1 point to any student selecting either A or B for question 17.

3.1.3 Subscale Sheets These sheets follow the subkeys (if any are used) and indicate which items appear in each subscale. There need to be exactly as many subscale sheets as indicated in column F of the main key sheet.

There is one sheet for each subscale, with the subscale name in columns 1-17, and the word "KEY" in columns 18-20. The subscale's number (1 through whatever number is coded in column F on the main key) is put in column F on the subscale sheet. Finally, all items appearing in the subscale have option A blackened in on this sheet. This mark does not mean A is the correct answer for question - it simply indicates the item belongs in the subscale. All other items are ignored in the scoring of this subscale.

3.2 Student Answer Sheets (Only General Purpose NCS answer sheets--available at the bookstore--can be used with this system.) Have students code in their names in the name block on the answer sheet, and whatever identification information is necessary in the identification block. This information will appear on the printouts returned to the students and on those prepared for the instructor.

Answer sheets can be re-used. Simply make sure that each student receives the same answer sheet each time, and prepare the main key, subkeys, and subscale cards for the items appearing on the test.

Consider this example: The first test has 50 questions (numbers 1-50 on the answer sheet) while the second has 75 questions (numbers 51-125 on the answer sheet). The first set of keys would be set up as usual, while the second ignores items 1-50 and considers only 51-125. This is done by (i) coding "75" in columns A-C on the main key

and (ii) blackening options for items 51-125 when indicating which point values are to be assigned to which options. Questions 1-50 are then ignored. The computer will skip items 1-50 in doing its analysis.

As a matter of policy, faculty should keep student answer sheets. Since the computer will print out an "echo" of each student's sheet (a page with the student's name and all the options he/she selected), it is not necessary to hand back answer sheets to the students.

Make sure the students are careful about both blackening boxes on the answer sheets appropriately and erasing carefully: The scanner will pick up extraneous marks on the answer sheets.

3.3 Submitting Answer Sheets to be Scored The completed set of materials (main key, subkeys if necessary, subscale sheets if necessary, and student answer sheets) should be submitted to be scanned at Upson II, Room 11-J. A brief "keypunch transmittal form" must be filled out there indicating the department, the instructor's name, and a phone number to be called when the scanning and scoring has been completed. THE TELEPHONE NUMBER IS CRITICAL: THIS NUMBER WILL BE CALLED ONCE WHEN THE TESTS HAVE BEEN SCORED. Consequently, it is best to have a departmental number so a message can be taken.

The person dropping off the tests to be scored needs to sign the transmittal form, and a signature is also required when the test results are picked up.

3.4 Printout and Its Interpretation The Generalized Item Analysis Program produces six different sets of information: Scoring key information, histograms (distributions of test scores), a record book sheet (a list of students' names and their test scores), answer sheet echos (transcripts of the answers selected by each student), item analysis information (item-by-item results), and total test statistics (means, standard deviations, etc.).

Scoring key information tells how many points were awarded for each option for each item, and the subscales (if any) which contained the item. It also provides an additional listing of all the items in each subscale. An example from the printout is:

```
PARAMETER LIST FOR:  SURVEY COURSE FIRST TEST
NUMBER OF ITEM IN TEST:  50
TEST TYPE IS NRT
OPTION VALUES FOR EACH ITEM IN TEST AND ALL SUBSCALES IN WHICH ITEM APPEARS
SUBSCALE 0 REPRESENTS DELETED ITEMS
(cont.)
```

ITEM	WEIGHT					SUBSCALES
	A	B	C	D	E	
1		1				1
2				1		1
3		1				2
4	1					2
5		1				1 2
6			1			2
.						
.						
.						
49				1		3
50		1				2

SUBSCALE NUMBER	NAME	MAXIMUM POINTS	ITEMS INCLUDED
	MAIN TEST	50	1 2 . . . 50
1	DEFINITIONAL ITEMS	5	1 2 5 8 11
2	BASIC CONCEPTS	25	3 4 5 6 . . . 50
.			
.			
.			
5	APPLYING CONCEPT	10	18 25 . . . 48

In this example, the instructor indicated the name of the test was "SURVEY COURSE FIRST EXAM", and that it was a norm-referenced test 50 items in length. One point was awarded to anyone selecting option B for item 1, 0 points for other options, for example, and this question appeared in subscale 1 only. In contrast, item 5 appeared in both subscales 1 and 2.

The items appearing in each subscale were also identified (i.e., The first subscale contained items 1, 2, 5, 8, and 11).

The importance of this information is that it (i) summarizes the makeup of the test, and (ii) it allows the instructor to check to make sure items were not miskeyed, that subscales contained the correct items, etc. Indeed, this is the first thing the instructor should inspect when the printouts are returned.

The second part of the printout contains the histograms (frequency distributions) of scores on the test. This is interpreted in a conventional manner (i.e., the higher the bar, the more students received the score indicated), and needs no special explanation.

The third part of the printout is the grade sheet. A sample grade sheet might look like this:

NAME	ID #	SUBSCALE NAME	RAW SCORE	SCALED SCORE
CLIFFORD T	0	MAIN TEST	40	592
		1 DEFINITIONAL ITEMS	3	500
		2 BASIC CONCEPTS	20	650
		3 . . .	.	.
		4 . . .	.	.
		5 APPLYING CONCEPTS	8	510
CLARK A	0	MAIN TEST	33	453
		1 DEFINITIONAL ITEMS	3	500
		2 BASIC CONCEPTS	17	608
		3 . . .	.	.
		4 . . .	.	.
		5 APPLYING CONCEPTS	7	485
JOHNSON A W	0	MAIN TEST	42	638
		1 DEFINITIONAL ITEMS	4	600
		2 BASIC CONCEPTS	20	650
		3 . . .	.	.
		4 . . .	.	.
		5 APPLYING CONCEPTS	9	570

The names of students appear in the left most column, and students' ID numbers follow. Since this teacher did not require students to fill in the ID box on the answer sheet, everyone's ID number is "0"; had ID's been filled in, these numbers would have appeared on the grade sheet. (Note that ID numbers can be assigned by the instructor; any unique number for each student will do.)

The subscale name refers to the part of the test being considered, and the raw score is the number of points the student has earned. T. Clifford, for example, had an overall score of 40 points, with 3 coming from the definitional items subscale, 20 from the basic concepts subscale, and so on. Clifford's status relative to the body of skill and knowledge covered on the test is reported as a scaled score (and is described under the heading of NRT's, section 4.3). Status is also reported for CRT's (highest criterion met) and DRT's (percent of domain mastered), and are each described under their respective headings.

The next set of pages in the printout are the echos of the answer sheets: For each question, they present the item number, the option selected by the student, the option awarded the greatest number of points, and the student's scores on the subscales and the total test. These sheets can be distributed to the students since they show both what the student did, and what constituted maximum performance.

Item analysis information follows immediately. These results indicate how many students selected each option as well as the mean and standard deviations of points awarded for each item. Clearly, the higher the mean, the better the performance of students on that item.

The discrimination index is a correlation coefficient indicating the relationship between performance on that item and the test as a whole. The discrimination index is described in more detail in sections 4.3 (NRT's), 5.3 (CRT's), and 6.3 (DRT's).

Finally, summary statistical information for each test is printed on the last page of the output. This information indicates the mean, standard deviation of test scores, and two measures of reliability: an internal consistency measure based on inter-item correlations, and a standard error or measurement based on the reliability coefficient. These measures are discussed in introductory text on psychometrics.

4.0 Norm-Referenced Tests The purpose of a norm-referenced test is to indicate how much each student knows relative to all other students in the class.* Questions appearing on NRT's are designed to discriminate among students of varying abilities; relatively small differences and scores for different students on well-constructed NRT's are indicative of meaningful differences in terms of those students' abilities. In order to put together a good NRT, the faculty member writes a series of test questions sharing a particular attribute: If a student responds correctly to a question, that student should also respond well to all other test questions, and conversely, any student missing a given question should do poorly on the test as a whole. If test questions of this variety are written and collected into a single exam, the test will have a capability of making relatively fine discriminations among students in the class.

NRT's scores are reported in a manner indicative of the student's place in the distribution. Centiles do this well, as do scaled scores such as those reported for the Scholastic Aptitude Tests prepared by the College Boards. These scaled scores have a mean of 500 and a standard deviation of 100 - approximately 68% of students have scores ranging from 400 (one standard deviation below the mean) to 600 (one standard deviation above the mean). Indeed, this scoring system is used to report performances on NRT's by this test scoring program: Two scores are reported for each student - a raw score (the raw number of points received by the student) and a scaled score.

4.1 Good and Poor Uses of NRT's NRT's are most useful in competitive situations such as deciding who is the most capable student among those being tested. In classroom settings, NRT's are used as a basis for grading on the curve; the distribution of NRT

scores is examined and decisions made on how to award grades follow "break points" in the distribution. There is a considerable debate among educators as to whether norm-referenced tests are useful for awarding grades in instructional settings. The issue concerns whether knowing more or less than one's colleagues constitutes having sufficient knowledge or competency. This issue will be addressed further under criterion-referenced tests (5.0).

4.2 Preparing Scoring Keys The test's main key must have the letters "NRT" in columns 15-17 (if these columns are left blank, the test is assumed to be an NRT), and the work "KEY" must appear in columns 18-20. The number of test questions may optionally be coded in columns A-C and the number of subscales (if appropriate), different point values (if appropriate), and the number of subscales (if appropriate) must also be coded in. All other instructions listed in section 3.1 should be followed.

4.3 Interpreting Results Since norm-referenced tests are designed to be discriminating, test questions should be reviewed to ascertain that they are working in this manner. One statistic listed in the printout is especially useful in this regard, and it is called the discrimination index: The correlation of performance on that question with performance on the test as a whole or the subscales in which the item appears. High discrimination indices indicate good NRT items, while low indices indicate deficient items. The minimum discrimination index value test constructors will consider to be satisfactory varies, and is a matter of the test constructor's experience, the nature of the group being examined (sophisticated students provide different test results than naive students) and the purpose to which the test is designed to meet. In general, discrimination indices of .4 or higher are cause for elation, while indices below .2 indicate items which probably ought not to be reused.

Items also ought to be reviewed according to the number of persons picking each option. The number of persons selecting each option (A-E) frequently provides information on why students perform as they do. For instance, when a particular option attracts little student attention, that option should be reviewed to determine whether or not it is indeed plausible from a student perspective: Sometimes the manner in which options are worded causes them to be more attractive or less attractive than originally intended, and this is reflected in the number of persons selecting it on the test. As a rule of thumb, incorrect answers in NRT's should be eliminated if they attract very little student attention (otherwise, they contribute nothing to the test). In short, one expends effort in reviewing both the discrimination indices and a pattern of responses in reviewing norm-referenced tests.

It is also important to examine the reliability coefficients for norm-referenced tests. These statistics, found near the end of the printout, indicate how internally consistent the test was. This is another way of estimating the degree to which the test discriminates - a test with a high reliability coefficient can make consistent discriminations among students of different ability more readily than a test with a lower reliability. As a rule of thumb, a reliability of .8 means a teacher can place confidence in decisions made about the entire class, while a reliability of .9 is required to make decisions about individual students. However, even though a high reliability is very desirable, when teachers must decide on grades decisions need to be made even in its absence.

Further, when there are relatively small numbers of students in comparison to a number of questions on the test (when the number of questions is less than half the number of students, say), test reliabilities are spuriously high for statistical reasons. Since this situation frequently applies in college classrooms (e.g., where there are 30 students and a 100 question test, for example), reliability is extremely difficult to estimate.

5.0 Criterion-Referenced Tests A criterion-referenced test is one which indicates whether a student's knowledge and/or skill exceeds a specified criterion level of performance. Typically, questions for these tests are selected on the basis of face validity: The test question clearly and directly measures what it is supposed to measure. Further, the content measured by the test question is an important aspect of what students are to learn, and not an incidental or relatively unimportant piece of information. Grades awarded indicate whether a student has reached a present criterion level of performance, and a student's score on the test is interpreted to mean the student either has a degree of mastery indicated by the criterion (e.g., in a pass/fail system, passing means "minimum essential"), and the student is said to have met or failed to meet the criterion. Note that multiple criteria can be used (up to five using this test scoring routine), and the usual letter grades (A through F) thus can be awarded according to which criteria the student has succeeded in meeting.

5.1 Good and Poor Uses of CRT's In most instructional settings, the professor's most important concern is whether students have sufficient knowledge or skill. In lower division courses, sufficiency may mean the student has learned enough to succeed in junior and senior level courses. A test score higher than others (as is indicated by a high score on NRT) is not necessarily evidence that the student has basic skills, concepts and principles well mastered since discriminating questions do not necessarily cover "basic" skill

and knowledge. Success in higher division courses would be more likely if the student took a criterion-reference test based on necessary skills, concepts, and principles and had a score exceeding the criterion established.

Some faculty believe that a system like this fails to be motivating to students who are competitive in nature. This circumstance can be addressed nicely by having multiple criteria for scoring tests. The lowest criterion would separate students with insufficient mastery from those demonstrating sufficient knowledge and skill while a more stringent criterion might be used to be assigned grades of A and B. It is critical to note that establishing a minimum criterion is not the same as awarding the grade of "D" in a norm-referenced system: A "D" simply means the student is at the bottom of distribution and may have only a spotty mastery of required knowledge and skill while a grade of "Pass" in a criterion-referenced setting means that all necessary knowledge and skill has been masterd. This important distinction is frequently not well understood.

5.2 Preparing Scoring Keys The letters "CRT" should be filled in in columns 15-17 on the main key sheet along with the word "KEY" in columns 18-20. Further, a number needs to be logged into column E indicating the number of criteria to be used in scoring students papers. Then, the worksheet attached to the end of this section of the manual needs to be duplicated and filled out in preparation for submitting the test in Room 11-J; these sheets indicate the names of the criteria to be used and the minimum number of points required for a student to meet each. Note that these criteria apply to the total test and not each of the subscale.

5.3 Interpreting Results The discrimination indices for questions on a criterion reference test represent the correlations of performance on each item with meeting or failing to meet each of the criteria specified by the faculty member. Under these circumstances, a high discrimination index identifies an item which was well understood by people reaching the criterion and poorly understood by those students failing to reach criterion. From an instructional point of view, such items represent pivotal aspects of instruction: A faculty member's attending these issues will allow students who fail to reach criterion to now perform beyond the specified level of competency.

Low discrimination indices do not necessarily indicate poor items in a criterion-referenced test. Because test questions measure only important issues, the fact that everyone responds correctly does not mean that the item is a poor one; indeed if 100% of the class responds correctly (and the item does not have any obvious errors in it), the faculty member can be satisfied that the entire class had mastery of the knowledge

or skill measured by the test question. Indeed a high mean score over all students on a good criterion-referenced test indicates that everyone did well on the non-trivial, important questions appearing on the test.

Reliability indices for CRT's are very difficult to interpret because conventional definitions of reliability are predicated on variation in performance among students, and variation is not necessarily an attribute of a criterion-referenced test.

6.0 Domain-Referenced Tests Domain-referenced tests are made up of test questions randomly sampled from a larger pool of items which operationally defines the domain of knowledge and skill students are expected to have mastered. The theory, DRT's are based on random sampling: A student's performance on a sample of items representing the entire domain informs on student's ability to answer all items in that domain. Consequently, a student's score is reported in terms of the percent of items responded to correctly. This percentage is then understood as an estimate of how the students would perform if he/she attempted all the items in the pool.

The chief problem with domain-referenced tests then, becomes preparing the item pool since large numbers of questions are needed to exhaust any given instructional topic or area. Faculty using domain-referenced tests, then, prepare test questions over a period of years, and randomly sample items from the maturing pool in order to construct tests for their students. Frequently, faculty members divide the pool into homogeneous strata and randomly sample items from each stratum for inclusion on the test. This precludes the test's being top heavy in one area and deficient in another; stratified random sampling serves to assure that a well-balanced test is produced.

6.1 Good and Poor Uses of DRT's DRT's are instructionally useful because they estimate the percent of the domain understood by students. This is especially important when the most desirable circumstances would be to test students on "everything"; judicious use of DRT's provides information that can be used in exactly the same way as information collected if all test questions had been asked. The deficiency of DRT's is that important questions may be skipped simply due to sampling error.

6.2 Preparing Scoring Keys Columns 15-17 on the main key need to include the designation DRT. Beyond that, the main key should be filled out as described in section 3.1.

6.3 Interpreting Results Discrimination indices for domain-referenced tests indicate the correlation of performance on the item with performance on all items on the test. High discrimination indices then suggest that (since all items are a random sampling from a pool) the item correlates well with hypothetical performance on the entire

domain. Thus, a domain-referenced test made up of items all of which have high discrimination indices provides a good estimate of how much a student really knows or can do. Note, however, that good estimates do not require large numbers of items with high indices; the more questions in the test, the better the estimate of the students' mean performances.

Patterns of response (the numbers of students picking each option A-E) are interpreted in precisely the same way as with NRT's and CRT's criterion-referenced tests.

7.0 Addenda

7.1 Glossary

Answer Key The answer key is a series of sheets which contain information the computer needs in order to score each test. The information contained on the answer key is found in section 3.1 and then in section 4.1 for NRT's, 5.1 CRT's and 6.1 DRT's.

CRT See criterion-referenced test.

Criterion-Referenced Test (CRT) This is a test where students' scores are compared to pre-existing criterion to determine the grades the students receive. CRT's are described in greater detail in section 5.0.

DRT See domain-referenced test.

Default When the user must identify one of a variety ofpossibilities (e.g., indicating whether the test is a CRT, DRT, or NRT), the computer assumes a default if nothing is specified on the key. In this example, the default is NRT, and failing to enter the test type in columns 15-17 will result in the exam being scored as an NRT.

Domain A test's domain is the skill and knowledge students must have in order to perform well on the examination. Domains can be defined as broadly or as narrowly as the test constructor requires them to be.

Domain-Referenced Test (DRT) A domain-referenced test (DRT) reports each student's score according to the percent of the domain the student is believed to have mastered. DRT's are described in detail in sections 1.1 and 6.0 of this manual.

Forced-Choice Item Any test question which provides students with a series of options to select among is a forced-choice item. These test questions differ from open-ended items in that students are expected to recognize rather than provide an answer.

Grade A grade is an indication of the student's status relative to the domain covered by the test. In the case of a norm-referenced test, the grade is reflective of where in the class' distribution the student falls; in a criterion-referenced test, the grade indicates the highest criterion the student has met, while in a domain-referenced test, the grade is simply the student's score - the percent of the domain the student is believed to have mastered.

Item An item is a test question. In the case of questions by this program, all items must be forced-choice in nature.

NRT See norm-referenced test.

Norm-Referenced Test (NRT) An NRT is a test which reports a students status by indicating whether the students performance was better or worse than that of other students. Norm-referenced tests are described in section 1.1 and 4.0 of this manual.

Test A test is a mechanism for finding out a student's status relative to the domain covered by the examination. Tests scored by this computer program are made up of a number of test questions which students must respond to. Other kinds of tests include open-ended examinations (where students may fill in blanks, provide short written answers or extended essays), or simulations.

Scaled Score A scaled score is a standardized way of reporting a students test performance relative to that of other students. In this computer program, scaled scores are used exclusively with norm-referenced tests. See section 4.3.

Status Status means the degree of competency a student has with the domain covered by the test. Depending on whether the test is an NRT, a DRT or a CRT, status is reported in one of the variety of manners. See sections 4.3, 5.3, and 6.3 for interpretation of status for each of the three kinds of tests.

Subkey This is a sheet prepared for the computer which indicates point values for each of the questions appearing on the tests. If there are four different point values to be awarded for each question (e.g., 5 points, 4 points, 3 points, 2 points) there will be three different subkey cards - each one indicating those options receiving a particular point value. See section 3.1 for a discussion of subkey cards.

Subscale A subscale is a set of test questions on measuring a related topic. This computer program will provide a score for each student on each subscale identified by the programs used here.

Subtests See subscale.

Test Questions See item.

Test Score The test score for a given student is the total number of points a student has received on the test. The test score is then used for awarding grades.

7.2 Examples

Example 1. A Norm-Referenced Test. The professor has indicated that the examination has 50 questions ("050" in columns A-C), and that each correct response is worth 1 point (the "1" in column H indicates only one possible point value, and the "01" in columns I-J means that value is 1).

The only other noteworthy aspects of the following key is NRTKEY at the end of the name box (telling the computer this is the key for a norm-referenced test). There are no subscales on this test (column F was left blank), and each of the questions numbered 1-50 will be scored for each student.

Example 2. A Domain-Referenced Test. This professor has produced a 100-item test to be scored as a DRT. The test designation appears at the end of the name box ("DRTKEY").

There are two possible point values for each question, and this fact is indicated by the "2" in column H. The options receiving 2 points (the value in column I-J) are those blackened in on the key. Notice that the next sheet has a "01" in columns I-J and this value will be awarded to students selecting the options blackened in for questions on that sheet. In other words, a student picking C for question 5 gets 2 points while someone picking B gets 1. No points are awarded for picking A, D, or E.

It is also important to note that questions on this test are divided into two subtests (also call subscales). The third sheet (labelled "DEFINITIONS") has a "1" in column F indicating that this is the first subscale. All the items with blackened "A's" on that sheet appear in the DEFINITIONS subscale. Similarly, the next sheet, named "PROBLEMS," is the second subscale (notice the "2" in column F), and it is made up of items 4-13, 16-17, etc.

Notice that the word "KEY" appears on all the sheets.

Example 3. Criterion-Referenced Test. This exam is designated as a criterion-referenced test with the characters "CRTKEY" in the name box. It has 40 questions ("040" in columns A-C), has two subscales, and only one point value--5 points are awarded for each correct response (see columns I-J).

A special sheet has been added to indicate how many points students need to reach the two criteria ("2" in column E). "Honors" will be awarded to anyone receiving 195 points, and "pass" to anyone receiving 185. Other information on this sheet is self-explanatory.

SIDE 1

GENERAL PURPOSE - NCS - ANSWER SHEET

FOR USE WITH ALL NCS SENTRY OPTICAL MARK READING SYSTEMS
SEE IMPORTANT MARKING INSTRUCTIONS ON SIDE 2

NAME (Last, First, M.I.)

MIDTERM EXAM NRTKEY

SEX (M) (F)

GRADE OR EDUC 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

BIRTH DATE			IDENTIFICATION NUMBER										SPECIAL CODES					
MO.	DAY	YR.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
			0	5	0					1	0	1						

Jan. Feb. Mar. Apr. May Jun. Jul. Aug. Sep. Oct. Nov. Dec.

EXAMPLE 1

21

22

1 A B C D E	11 A B C D E	21 A B C D E	31 A B C D E	41 A B C D E
2 A B C D E	12 A B C D E	22 A B C D E	32 A B C D E	42 A B C D E
3 A B C D E	13 A B C D E	23 A B C D E	33 A B C D E	43 A B C D E
4 A B C D E	14 A B C D E	24 A B C D E	34 A B C D E	44 A B C D E
5 A B C D E	15 A B C D E	25 A B C D E	35 A B C D E	45 A B C D E
6 A B C D E	16 A B C D E	26 A B C D E	36 A B C D E	46 A B C D E
7 A B C D E	17 A B C D E	27 A B C D E	37 A B C D E	47 A B C D E
8 A B C D E	18 A B C D E	28 A B C D E	38 A B C D E	48 A B C D E
9 A B C D E	19 A B C D E	29 A B C D E	39 A B C D E	49 A B C D E
10 A B C D E	20 A B C D E	30 A B C D E	40 A B C D E	50 A B C D E
51 A B C D E	61 A B C D E	71 A B C D E	81 A B C D E	91 A B C D E
52 A B C D E	62 A B C D E	72 A B C D E	82 A B C D E	92 A B C D E
53 A B C D E	63 A B C D E	73 A B C D E	83 A B C D E	93 A B C D E
54 A B C D E	64 A B C D E	74 A B C D E	84 A B C D E	94 A B C D E
55 A B C D E	65 A B C D E	75 A B C D E	85 A B C D E	95 A B C D E
56 A B C D E	66 A B C D E	76 A B C D E	86 A B C D E	96 A B C D E
57 A B C D E	67 A B C D E	77 A B C D E	87 A B C D E	97 A B C D E
58 A B C D E	68 A B C D E	78 A B C D E	88 A B C D E	98 A B C D E
59 A B C D E	69 A B C D E	79 A B C D E	89 A B C D E	99 A B C D E
60 A B C D E	70 A B C D E	80 A B C D E	90 A B C D E	100 A B C D E

TM

SIDE 1

GENERAL PURPOSE - NCS - ANSWER SHEET

FOR USE WITH ALL NCS SENTRY OPTICAL MARK READING SYSTEMS
SEE IMPORTANT MARKING INSTRUCTIONS ON SIDE 2

NAME (Last, First, M.I.)

FINAL EXAM DRTKEY

SEX (M) (F)

GRADE OR EDUC 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

BIRTH DATE			IDENTIFICATION NUMBER										SPECIAL CODES					
MO.	DAY	YR.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
			1	0	0			2		2	0	2						
Jan.																		
Feb.																		
Mar.	0	0	0															
Apr.	1	1	1															
May	2	2	2															
Jun.	3	3	3															
Jul.		4	4															
Aug.		5	5															
Sep.		6	6															
Oct.		7	7															
Nov.		8	8															
Dec.		9	9															

EXAMPLE 2 - Sheet 1

	A	B	C	D	E
1		●			
2	●				
3				●	
4					●
5			●		
6	●				
7				●	
8		●			
9			●		
10					●
11					●
12			●		
13			●		
14	●				
15		●			
16		●			
17					●
18				●	
19	●				
20				●	
21		●			
22			●		
23	●				
24				●	
25		●			
26					●
27			●		
28	●				
29					●
30				●	
31				●	
32			●		
33					●
34	●				
35				●	
36			●		
37	●				
38		●			
39					●
40		●			
41			●		
42		●			
43				●	
44			●		
45	●				
46		●			
47				●	
48					●
49	●				
50					●
51				●	
52	●				
53			●		
54	●				
55			●		
56		●			
57					●
58	●				
59		●			
60					●
61			●		
62	●				
63				●	
64			●		
65				●	
66		●			
67		●			
68					●
69	●				
70					●
71		●			
72			●		
73				●	
74		●			
75	●				
76			●		
77				●	
78	●				
79					●
80					●
81				●	
82			●		
83				●	
84					●
85		●			
86	●				
87			●		
88					●
89		●			
90			●		
91		●			
92	●				
93	●				
94					●
95			●		
96				●	
97		●			
98				●	
99			●		
100					●

TM

23

24

SIDE 1

GENERAL PURPOSE - NCS - ANSWER SHEET

FOR USE WITH ALL NCS SENTRY OPTICAL MARK READING SYSTEMS
SEE IMPORTANT MARKING INSTRUCTIONS ON SIDE 2

NAME (Last, First, M.I.) KEY

SEX (M) (F)

GRADE OR EDUC 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14

BIRTH DATE			IDENTIFICATION NUMBER										SPECIAL CODES					
MO.	DAY	YR.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Jan.											0	1						
Feb.																		
Mar.	0	0	0	0	0	0	0	0	0	0	●	0	0	0	0	0	0	0
Apr.	1	1	1	1	1	1	1	1	1	1	1	●	1	1	1	1	1	1
May	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Jun.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Jul.		4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Aug.		5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Sep.		6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
Oct.		7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
Nov.		8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
Dec.		9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

EXAMPLE 2 - Sheet 2

	A	B	C	D	E
1	①	②	③	④	⑤
2	①	②	③	④	⑤
3	①	②	③	④	⑤
4	①	②	③	④	⑤
5	①	●	③	④	⑤
6	①	②	③	④	⑤
7	①	②	③	④	⑤
8	①	②	③	④	⑤
9	①	②	③	④	⑤
10	①	②	③	④	⑤

	A	B	C	D	E
11	①	②	③	④	⑤
12	①	②	③	④	⑤
13	①	②	③	④	⑤
14	①	②	③	④	⑤
15	①	②	③	④	⑤
16	①	②	③	④	⑤
17	①	●	③	④	⑤
18	①	②	③	④	⑤
19	①	②	③	④	⑤
20	①	②	③	④	⑤

	A	B	C	D	E
21	①	②	③	④	⑤
22	①	②	③	④	⑤
23	①	②	③	④	⑤
24	①	②	③	④	⑤
25	①	②	③	④	⑤
26	①	②	③	④	⑤
27	①	②	③	●	⑤
28	①	②	③	④	⑤
29	①	②	③	④	⑤
30	①	②	③	④	⑤

	A	B	C	D	E
31	①	②	③	④	⑤
32	①	②	③	④	⑤
33	①	②	③	④	⑤
34	①	②	③	●	⑤
35	①	②	③	④	⑤
36	①	②	③	④	⑤
37	①	②	③	④	⑤
38	①	②	③	④	⑤
39	①	②	③	④	⑤
40	①	②	③	④	⑤

	A	B	C	D	E
41	①	②	③	④	⑤
42	①	②	③	④	⑤
43	①	②	③	④	⑤
44	①	②	③	④	⑤
45	①	②	③	④	⑤
46	①	②	③	④	⑤
47	①	②	③	④	⑤
48	①	②	③	④	⑤
49	①	②	③	④	⑤
50	①	②	③	④	⑤

	A	B	C	D	E
51	①	②	③	④	⑤
52	①	②	③	④	⑤
53	①	②	③	④	⑤
54	①	②	③	④	⑤
55	①	②	③	④	⑤
56	①	②	③	④	⑤
57	①	②	③	④	⑤
58	①	②	③	④	⑤
59	①	②	③	④	⑤
60	①	②	③	④	⑤

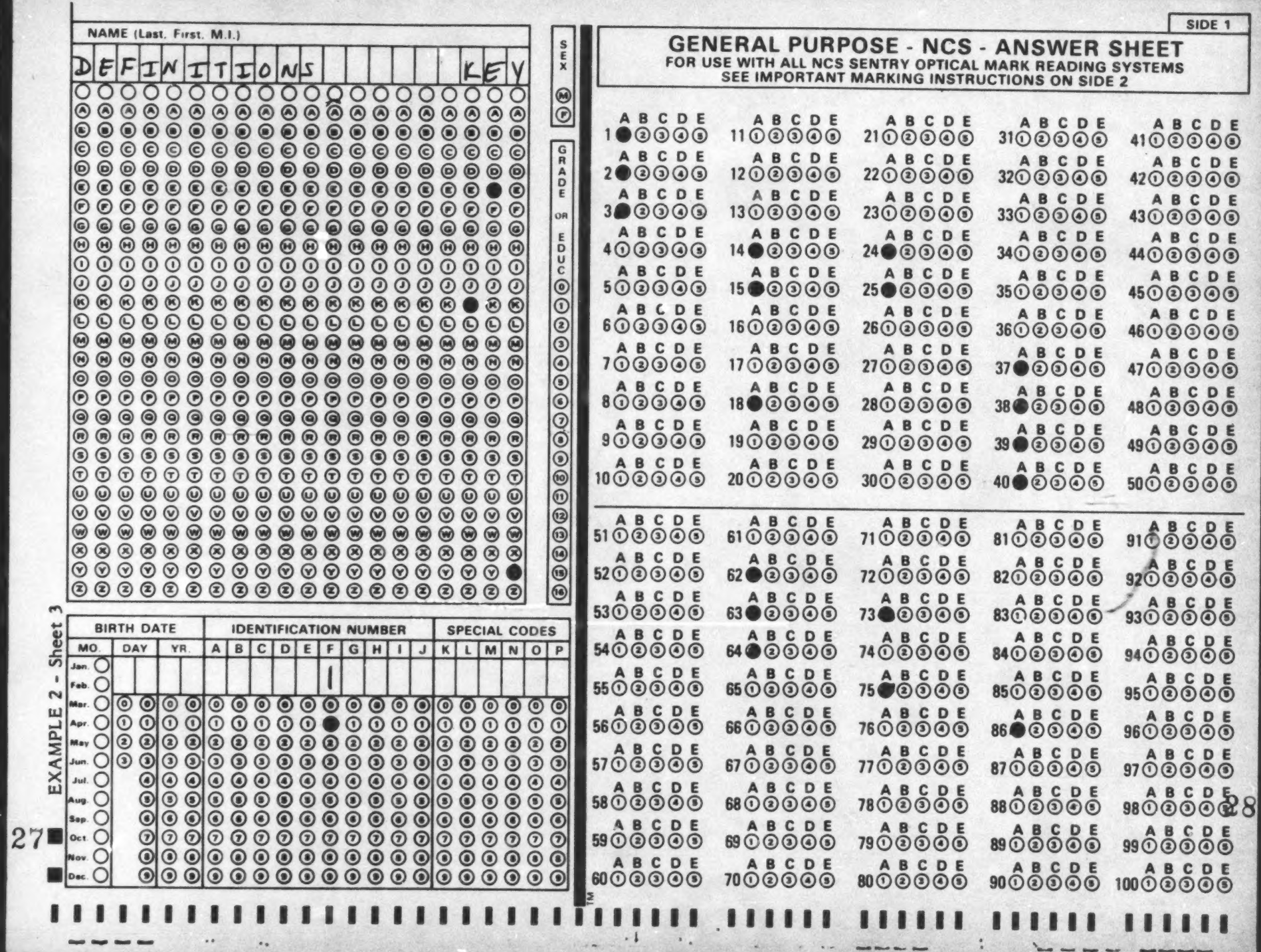
	A	B	C	D	E
61	①	②	③	④	⑤
62	①	②	③	④	⑤
63	①	②	③	④	⑤
64	①	②	③	④	⑤
65	①	②	③	④	⑤
66	①	②	③	④	⑤
67	①	②	③	④	⑤
68	①	②	③	④	⑤
69	①	②	③	④	⑤
70	①	②	③	④	⑤

	A	B	C	D	E
71	①	②	③	④	⑤
72	①	②	③	④	⑤
73	①	②	③	④	⑤
74	①	②	③	④	⑤
75	①	②	③	④	⑤
76	①	②	③	④	⑤
77	①	②	③	④	⑤
78	①	②	③	④	⑤
79	①	②	③	④	⑤
80	①	②	③	④	⑤

	A	B	C	D	E
81	①	②	③	④	⑤
82	①	②	③	④	⑤
83	①	②	③	④	●
84	①	②	③	④	⑤
85	①	②	③	④	⑤
86	①	②	③	④	⑤
87	①	②	③	④	⑤
88	①	②	③	④	⑤
89	①	②	③	④	⑤
90	①	②	③	④	⑤

	A	B	C	D	E
91	①	②	③	④	⑤
92	①	②	③	④	⑤
93	①	②	③	④	⑤
94	①	②	③	④	⑤
95	○	②	③	④	●
96	①	②	③	④	⑤
97	①	②	③	④	⑤
98	①	②	③	④	⑤
99	①	②	③	④	⑤
100	①	②	③	④	⑤

TM


SIDE 1
GENERAL PURPOSE - NCS - ANSWER SHEET
FOR USE WITH ALL NCS SENTRY OPTICAL MARK READING SYSTEMS
SEE IMPORTANT MARKING INSTRUCTIONS ON SIDE 2
NAME (Last, First, M.I.)
DEFINITIONS KEY
SEX M F
GRADE OR EDUC
BIRTH DATE
MO. DAY YR.
IDENTIFICATION NUMBER
SPECIAL CODES
EXAMPLE 2 - Sheet 3
27
28

SIDE 1

GENERAL PURPOSE - NCS - ANSWER SHEET

FOR USE WITH ALL NCS SENTRY OPTICAL MARK READING SYSTEMS
SEE IMPORTANT MARKING INSTRUCTIONS ON SIDE 2

NAME (Last, First, M.I.)

PROBLEMS KEY

SEX (M) (F)

GRADE OR EDUC 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14

BIRTH DATE			IDENTIFICATION NUMBER										SPECIAL CODES					
MO.	DAY	YR.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Jan.								2										
Feb.																		
Mar.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Apr.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
May	2	2	2	2	2	2	2	●	2	2	2	2	2	2	2	2	2	2
Jun.	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Jul.		4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Aug.		5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Sep.		6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
Oct.		7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
Nov.		8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
Dec.		9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

EXAMPLE 2 - Sheet 4

1 A B C D E	11 A B C D E	21 A B C D E	31 A B C D E	41 A B C D E
2 A B C D E	12 A B C D E	22 A B C D E	32 A B C D E	42 A B C D E
3 A B C D E	13 A B C D E	23 A B C D E	33 A B C D E	43 A B C D E
4 A B C D E	14 A B C D E	24 A B C D E	34 A B C D E	44 A B C D E
5 A B C D E	15 A B C D E	25 A B C D E	35 A B C D E	45 A B C D E
6 A B C D E	16 A B C D E	26 A B C D E	36 A B C D E	46 A B C D E
7 A B C D E	17 A B C D E	27 A B C D E	37 A B C D E	47 A B C D E
8 A B C D E	18 A B C D E	28 A B C D E	38 A B C D E	48 A B C D E
9 A B C D E	19 A B C D E	29 A B C D E	39 A B C D E	49 A B C D E
10 A B C D E	20 A B C D E	30 A B C D E	40 A B C D E	50 A B C D E
51 A B C D E	61 A B C D E	71 A B C D E	81 A B C D E	91 A B C D E
52 A B C D E	62 A B C D E	72 A B C D E	82 A B C D E	92 A B C D E
53 A B C D E	63 A B C D E	73 A B C D E	83 A B C D E	93 A B C D E
54 A B C D E	64 A B C D E	74 A B C D E	84 A B C D E	94 A B C D E
55 A B C D E	65 A B C D E	75 A B C D E	85 A B C D E	95 A B C D E
56 A B C D E	66 A B C D E	76 A B C D E	86 A B C D E	96 A B C D E
57 A B C D E	67 A B C D E	77 A B C D E	87 A B C D E	97 A B C D E
58 A B C D E	68 A B C D E	78 A B C D E	88 A B C D E	98 A B C D E
59 A B C D E	69 A B C D E	79 A B C D E	89 A B C D E	99 A B C D E
60 A B C D E	70 A B C D E	80 A B C D E	90 A B C D E	100 A B C D E

TM

25

30

SIDE 1

GENERAL PURPOSE - NCS - ANSWER SHEET

FOR USE WITH ALL NCS SENTRY OPTICAL MARK READING SYSTEMS
SEE IMPORTANT MARKING INSTRUCTIONS ON SIDE 2

NAME (Last, First, M.I.)

UNIT TEST CRTKEY

SEX (M) (F)

GRADE OR EDUC 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

BIRTH DATE				IDENTIFICATION NUMBER										SPECIAL CODES					
MO.		DAY		YR.		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Jan.																					
Feb.						0	4	0		2			1	0	5						
Mar.		0	0	0	0		0		0	0	0	0	0		0	0	0	0	0	0	0
Apr.		1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1
May		2	2	2	2	2	2	2	2		2	2	2	2	2	2	2	2	2	2	2
Jun.		3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Jul.			4	4	4	4		4	4	4	4	4	4	4	4	4	4	4	4	4	4
Aug.			5	5	5	5	5	5	5	5	5	5	5	5		5	5	5	5	5	5
Sep.			6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
Oct.			7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
Nov.			8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
Dec.			9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9

EXAMPLE 3

31

1	A B C D E	11	A B C D E	21	A B C D E	31	A B C D E	41	A B C D E
2	A B C D E	12	A B C D E	22	A B C D E	32	A B C D E	42	A B C D E
3	A B C D E	13	A B C D E	23	A B C D E	33	A B C D E	43	A B C D E
4	A B C D E	14	A B C D E	24	A B C D E	34	A B C D E	44	A B C D E
5	A B C D E	15	A B C D E	25	A B C D E	35	A B C D E	45	A B C D E
6	A B C D E	16	A B C D E	26	A B C D E	36	A B C D E	46	A B C D E
7	A B C D E	17	A B C D E	27	A B C D E	37	A B C D E	47	A B C D E
8	A B C D E	18	A B C D E	28	A B C D E	38	A B C D E	48	A B C D E
9	A B C D E	19	A B C D E	29	A B C D E	39	A B C D E	49	A B C D E
10	A B C D E	20	A B C D E	30	A B C D E	40	A B C D E	50	A B C D E
51	A B C D E	61	A B C D E	71	A B C D E	81	A B C D E	91	A B C D E
52	A B C D E	62	A B C D E	72	A B C D E	82	A B C D E	92	A B C D E
53	A B C D E	63	A B C D E	73	A B C D E	83	A B C D E	93	A B C D E
54	A B C D E	64	A B C D E	74	A B C D E	84	A B C D E	94	A B C D E
55	A B C D E	65	A B C D E	75	A B C D E	85	A B C D E	95	A B C D E
56	A B C D E	66	A B C D E	76	A B C D E	86	A B C D E	96	A B C D E
57	A B C D E	67	A B C D E	77	A B C D E	87	A B C D E	97	A B C D E
58	A B C D E	68	A B C D E	78	A B C D E	88	A B C D E	98	A B C D E
59	A B C D E	69	A B C D E	79	A B C D E	89	A B C D E	99	A B C D E
60	A B C D E	70	A B C D E	80	A B C D E	90	A B C D E	100	A B C D E

32

EXAMPLE 3 - Sheet 2

****
* CRITERION-REFERENCED TEST *
* SUPPLEMENTAL DATA WORKSHEET *
* FOR EXAM ANALYSIS *
****

Test name: UNIT TEST (from master key)

Submitted by: H. B. Slotnick

Date: 15 March, 1984

This form must be submitted along with the keys and student answer sheets for any exams which are to use the CRT analysis. The maximum number of criteria allowed is 5. Criterion names can be no longer than 10 characters each.

	Criterion Name	Lowest Allowable Score
1)	Honors	195
2)	Pass	185
3)		
4)		
5)		

Example:

1)	A	90
2)	B	80
3)	C	70
4)	D	60

etc.

7.3

PLANNING A TEST

1. DECIDE HOW TEST IS TO BE USED (PURPOSE)
2. DECIDE WHAT TO TEST (OUTCOMES)
   - A. ASPECTS OF ACHIEVEMENT
   - B. COMPLEX VS. EFFICIENT TASKS
   - C. CONTENT (IMPORTANT OUTCOMES)
3. DECIDE WHEN TO TEST (FREQUENCY AND TIMING)
4. DECIDE ON TYPE OF TEST AND TYPE OF ITEM (OUTCOME MEASURE)
   - A. ESSAY
     1. RESTRICTED RESPONSE
     2. EXTENDED RESPONSE
     3. ORAL
   - B. OBJECTIVE
     1. MULTIPLE-CHOICE
     2. TRUE-FALSE
     3. MATCHING
     4. CLASSIFICATION
     5. SHORT ANSWER
   - C. PROBLEM/PERFORMANCE
     1. CONTINUOUS OBSERVATION
     2. RATING SCALE
     3. CHECKLIST
     4. PEER APPRAISAL
5. DECIDE ON HOW LONG TEST SHOULD TAKE (NUMBER OF QUESTIONS AND/OR MINUTES)
6. DECIDE ON MEANS AND FORMAT FOR PRESENTING TEST (TESTING SITUATION)

For More Information or Assistance:
Office of Instructional Development
University of North Dakota
407 Twamley Hall (701)777-3325

7.4 Rules of Thumb: Essay, Objective & Performance Tests

Essay Tests

Rules for construction:

1. Use essay questions to measure complex learning outcomes only.
2. Relate the questions as directly as possible to the learning outcomes being measured.
3. Formulate questions which present a clear, definite task to the student.
4. Do not permit students a choice of questions unless the learning outcome requires it.
5. Provide ample time for answering and give suggested time limits for each question.

Rules for scoring:

1. Evaluate answers to essay questions in terms of the learning outcomes being measured.
2. Score restricted response answers by the point method, using a model answer as a guide.
3. Grade extended response answers by the rating method, using clearly defined criteria as a guide.
4. Evaluate all of the students' answers to one question before proceeding to the next question.
5. Evaluate students' answers to essay questions anonymously.
6. Have two or more persons grade each answer, wherever possible.

Objective Tests

Multiple choice items:

Rules for construction:

1. Design each item to measure an important learning outcome.
2. Present a single, clearly formulated problem in the stem of the item.
3. State the stem of the item in simple, clear language.
4. Put as much of the wording as possible in the stem of the item.
5. State the stem of the item in positive form, wherever possible.
6. Emphasize negative wording whenever it is used in the stem of an item.
7. Make certain that the intended answer is correct or clearly best.
8. Make all alternatives grammatically consistent with the stem of the item and parallel in form.
9. Avoid verbal clues which might enable students to select the correct answer or to eliminate an incorrect alternative.
   (a) Similarity of wording in both the stem and the correct answer is one of the more obvious clues. Key words in the stem may be unintentionally repeated verbatim in the correct answer, a synonym may be used, or the words may simply sound or look alike.
   (b) Stating the correct answer in textbook language or stereotyped phraseology may cause the student to select it because it looks better than the other alternatives, or because he vaguely recalls having seen it before.
   (c) Stating the correct answer in greater detail may provide a clue. Also, when the answer is qualified by modifiers which are typically associated with the true statements (e.g., sometimes, may, usually), it is more likely to be chosen.

(d) Including absolute terms in the distracters enables students to eliminate them as possible answers, because such terms are commonly associated with false statements (e.g., always, never, all, none, only). This makes the correct answer obvious or at least increases the chances of guessing it.
(e) Including two responses that are all-inclusive makes it possible to eliminate the other alternatives, since one of the two obviously must be the correct answer.
(f) Including two responses that have the same meaning makes it possible to eliminate them as potential answers. If two alternatives have the same meaning and only one answer is to be selected, it is fairly obvious that both of them must be incorrect.

10. Make the distracters plausible and attractive to the uninformed.
(a) Use the common misconceptions, or common errors, of students as distracters.
(b) State the alternatives in the language of the student.
(c) Use good sounding words (e.g., accurate, important, etc.) in the distracters, as well as in the correct answers.
(d) Make the distracters similar to the correct answer in both length and complexity of wording.
11. Vary the relative length of the correct answer to eliminate length as a clue.
12. Avoid use of the alternative "all of the above" and use "none of the above" with extreme caution.
13. Vary the position of the correct answer in a random manner.
14. Control the difficulty of the item either by varying the problem in the stem or by changing the alternatives.
15. Make certain each item is independent of the other items in the test.
16. Use an efficient item format.

True-False Items:

Rules for construction:

1. Include only one central significant idea in each statement.
2. Word the statement so precisely that it can be judged unequivocally true or false.
3. Keep the statements short and use simple language structure.
4. Use negative statements sparingly and avoid double negatives.
5. Statements of opinion should be attributed to some source.
6. Avoid extraneous clues to the answer.

Matching Items:

Rules for construction:

1. Include only homogeneous material in each matching item.
2. Keep the lists of items short and put the brief responses on the right.
3. Use a larger, or smaller, number of responses than premises, and permit the responses to be used more than once.
4. Specify in the directions the basis for matching and indicate that each response may be used once, more than once, or not at all.

Short Anser Items:

Rules for construction:

1. State the item so that only a single, brief answer is possible.
2. Start with a direct question and switch to an incomplete statement only when greater conciseness is possible.

3. The words to be supplied should relate to the main point of the statement.
4. Place the blanks at the end of the statement.
5. Avoid extraneous clues to the answer.
6. For numerical answers, indicate the degree of precision expected and the units in which they are to be expressed.

Problem/Performance Tests

Rules for construction and use:

1. Clearly define the skill(s) to be observed (these may be thinking skills, interpersonal skills or physical skills). These should match your course objectives.
2. Select a natural or simulated situation in which to observe the performance.
3. Specify task elements of greatest relevance to the quality of the performance, for example, speed of performance, accuracy, number and seriousness of procedural errors, errors in following instructions, selection of proper materials or techniques, coherence and appropriateness of sequence of performance, intensity and force of the performance.
   Select elements in proportion to emphasis given in the course.
   Judge whether these elements can be accurately and efficiently evaluated.
4. Problems in performance evaluation include a) tendency to use exclusively precise categories (especially with rating scales and checklists) and b) counter-tendency to use overly broad and vague categories (especially with anecdotal records and peer evaluation).
5. Common rating errors include a) being too lenient because of the difficulties of rating, b) rating everyone near the average performance and c) letting performance of one skill effect rating on other skills or elements of performance (the Halo effect).
6. In training other raters (e.g., clinical supervisors, lab assistants, peer raters) develop a common frame of performance, common set of criteria, and common procedures. Do this by a) practice sessions for raters, b) review and comparison among raters, c) computation of inter-rater reliabilities scores and d) replacement of inconsistent raters or raters who cannot work with common frame of reference.
7. Special situations for performance evaluation might include leaderless groups, problem-solving teams, individual and group role playing, actual performance situations.

Guidelines for Reviewing and Editing Items

1. Does each test item measure an important learning outcome included in the table of specifications?
2. Is each item type appropriate for the particular learning outcome to be measured?
3. Does each item present a clearly formulated task?
4. Is the item stated in simple, clear language?
5. Is the item free from extraneous clues?
6. Is the difficulty of the item appropriate for the students to be tested?

7. Is each test item independent and are the items, as a group, free from overlapping?
8. Do the items to be included in the test provide adequate coverage of the table of specifications?

Arranging Items

1. The items should be arranged so that all items of the same type are grouped together.
2. The items should be arranged in order of increasing difficulty.
3. For some purposes, it may be desirable to group together items which measure the same learning outcomes or the same subject matter content.

Ounce of Prevention Principle

One hour of colleague review of test is better than five hours of student complaints.

Colleagues can (1) look for congruence of test/items and course objectives, (2) look for ambiguity in phrasing and (3) review scoring key for accuracy and emphasis.

For More Information or Assistance:
Office of Instructional Development
University of North Dakota
407 Twamley Hall (701)777-3325